Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Model-assisted optimal allocation for planned domains using composite estimation

by Wilford B. Molefe and Robert Graham Clark

Release date: December 17, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2015

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 377-387
Statistics Canada, Catalogue No. 12-001-X

Model-assisted optimal allocation for planned domains using composite estimation

Wilford B. Molefe and Robert Graham Clark[1]

Abstract

This paper develops allocation methods for stratified sample surveys where composite small area estimators are a priority, and areas are used as strata. Longford (2006) proposed an objective criterion for this situation, based on a weighted combination of the mean squared errors of small area means and a grand mean. Here, we redefine this approach within a model-assisted framework, allowing regressor variables and a more natural interpretation of results using an intra-class correlation parameter. We also consider several uses of power allocation, and allow the placing of other constraints such as maximum relative root mean squared errors for stratum estimators. We find that a simple power allocation can perform very nearly as well as the optimal design even when the objective is to minimize Longford's (2006) criterion.

Key Words: Small area estimation; Sample design; Sample size allocation; Composite estimation; Mean squared error.

1 Introduction

Sample surveys have long been used as cost-effective means for data collection but it is also the case that general purpose surveys will often not achieve adequate precision for statistics for subpopulations of interest (often called domains or areas). Domains may be geographically based areas such as states. They may also be cross-classifications of a small geographic area and a specific demographic or social group. A domain is regarded as *small* if the domain-specific sample is not large enough to produce a direct estimator with satisfactory precision.

In this paper, we suppose that stratified sampling is used with H strata defined by the small areas, indexed by $h \in U^1$. The set of all N units in the population is denoted U and the set of H strata is denoted U^1. This effectively assumes that small areas can be identified in advance, which is not always the case (Marker 2001). Even so, the survey designer may be able to make an educated guess at areas of interest, which should still result in an improved design even if new requirements for area statistics emerge after the survey has been run. The population of N_h units in stratum h is written U_h and the sample of n_h units selected by simple random sampling without replacement (SRSWOR) from stratum h is s_h. Let Y_j be the value of the characteristic of interest for the j^{th} unit in the population. The small area population mean for stratum h is $\bar{Y}_h$ and the national mean is $\bar{Y}$. The corresponding sample estimators are $\bar{y}_h$ and $\bar{y}$, respectively; $\bar{y}_h = n_h^{-1}\sum_{j\in s_h} y_j$ and $\bar{y} = \sum_{h\in U^1} P_h \bar{y}_h$, where $P_h = N_h/N$. Let the sampling variances be $v_h = \text{var}_p(\bar{y}_h)$ and $v = \text{var}_p(\bar{y})$.

Longford (2006) considers the problem of optimal sample sizes for small area estimation for this design. The approach is based on minimizing the weighted sum of the mean squared errors of the planned small area mean estimators and an overall estimator of the mean. The weight attached to each area is proportional to the area population raised to the q^{th} power, so the value of $0 \le q \le 2$ specifies the

1. Wilford B. Molefe, Department of Statistics, University of Botswana. E-mail: molefewb@mopipi.ub.bw; Robert Graham Clark, National Institute for Applied Statistics Research Australia, University of Wollongong. E-mail: rclark@uow.edu.au.

relative importance of larger compared to smaller areas. The mean squared error of the all-strata mean estimator is multiplied by G, where G reflects the perceived priority of this estimator. An analytical solution exists for the case where $G = 0$, but it has undesirable practical properties, and may sometimes result in zero or minimum sample sizes for some strata. When $G > 0$, Longford (2006) suggests numerical optimization.

Choudhry, Rao and Hidiroglou (2012) investigate the use of nonlinear programming (NLP) to efficiently allocate sample to strata, when there may be bounds on stratum sample sizes, and priority on overall, stratum and cross-strata domain estimators of multiple variables. The paper mostly concentrates on design-based direct area estimators, but they also consider the objective criterion of Longford (2006) for composite estimation. For the Canadian Monthly Retail Trade Survey, they show that the Longford allocation gives extremely unequal sample sizes by strata, for q equal to 0.5, 1 and 1.5. For example, when $q = 1.5$, the highest stratum coefficient of variation (CV) is 112%, and even for $q = 0.5$, the highest coefficient of variation is 24%, which was deemed too high. It is not clear whether these CV%s refer to direct or composite estimators - such high CVs would be surprising for composite estimators, as their CVs are bounded above even as the sample size tends to zero. Choudhry et al. (2012) did not investigate whether other designs such as power allocations can give low values of Longford's criteria.

The aim of this paper is to find the best allocation to strata for a linear combination of the mean squared errors of small area composite estimators and of an overall estimator of the mean, similar to Longford (2006). In Section 2 we reformulate the objective in model-assisted terms, introduce the use of regressor variables, and derive a model-assisted composite estimator. Section 3 is devoted to optimizing the design. In Subsection 3.1 we discuss direct optimization, for example by NLP. Subsection 3.2 describes power allocation with the exponent chosen to numerically minimize the objective criterion. Section 4 is a numerical study of the various methods using the Swiss canton data of Longford (2006) and Section 5 contains conclusions.

2 Composite estimation

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $(1 - \phi_h)\,\bar{y}_h + \phi_h \bar{y}$ of the sample mean $\bar{y}_h$ for the target area h and the overall sample mean $\bar{y}$ of the target variable. The coefficients ϕ_h are set with the intent to minimise its mean squared error (MSE), see for example Rao (2003, Section 4.3). The coefficients by which the MSE is minimized depend on some unknown parameters which have to be estimated.

Better results can be obtained if there are some regressors $\mathbf{x}_i$, for which domain population means are available, as well as sample data at either unit or domain level enabling Y to be regressed on $\mathbf{x}$. A synthetic estimator for domain h is then defined by $\hat{Y}_{h(\mathrm{syn})} = \hat{\boldsymbol{\beta}}^{\mathrm{T}} \bar{\mathbf{X}}_h$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient, and $\bar{\mathbf{X}}_h$ is the domain population mean of the regressor variables. An efficient direct estimator which is particularly suitable when domain sizes may be small is $\bar{y}_{hr} = \bar{y}_h + \hat{\boldsymbol{\beta}}^{\mathrm{T}} (\bar{\mathbf{x}}_h - \bar{\mathbf{X}}_h)$ (Hidiroglou and Patak 2004) where $\bar{y}_h$ and $\bar{\mathbf{x}}_h$ are the domain h sample means of Y and X. A composite estimator can then be constructed as $\tilde{y}_h^C = (1 - \phi_h)\,\bar{y}_{hr} + \phi_h \hat{\boldsymbol{\beta}}^{\mathrm{T}} \bar{\mathbf{X}}_h$.

The design-based MSE of the composite estimator is given by:

$$\text{MSE}_p\left(\bar{y}_h^C; \bar{Y}_h\right) = (1-\phi_h)^2 v_{hr} + \phi_h^2 \left\{v_{h(\text{syn})} + B_h^2\right\} + 2\phi_h (1-\phi_h) c_h$$

where c_h is the sampling covariance of $\bar{y}_{hr}$ and $\hat{\bar{Y}}_{h(\text{syn})}$, v_{hr} is the sampling variance of the direct estimator $\bar{y}_{hr}$, $v_{h(\text{syn})}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{h(\text{syn})}$, and $B_h = \boldsymbol{\beta}_U^{\mathrm{T}} \bar{\mathbf{X}}_h - \bar{Y}_h$ is the bias of using $\hat{\bar{Y}}_{h(\text{syn})}$ to estimate $\bar{Y}_h$, with $\boldsymbol{\beta}_U$ denoting the approximate design-based expectation of $\hat{\boldsymbol{\beta}}$. Further,

$$\text{MSE}_p\left(\bar{y}_h^C; \bar{Y}_h\right) \approx (1-\phi_h)^2 v_{h(\text{syn})} + \phi_h^2 B_h^2 \tag{2.1}$$

because $c_h \ll v_h$ and $v \ll v_h$ when the number of small areas is large, under regularity conditions.

A two-level linear model ξ conditional on the values of $\mathbf{x}$ will be assumed, with uncorrelated stratum random effects u_h and unit residuals ε_i :

$$\left.\begin{aligned} Y_i &= \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i + u_h + \varepsilon_i \\ E_\xi[u_h] &= E_\xi[\varepsilon_i] = 0 \\ \text{var}_\xi[u_h] &= \sigma_{uh}^2 \\ \text{var}_\xi[\varepsilon_j] &= \sigma_{eh}^2 \end{aligned}\right\} \tag{2.2}$$

for $h \in U^1$ and $i \in U_h$. This implies that $\text{var}_\xi[Y_i] = \sigma_{uh}^2 + \sigma_{eh}^2 = \sigma_h^2$ for all $i \in U$, and that the covariance $\text{cov}_\xi[Y_i, Y_j]$ equals $\rho_h \sigma_h^2$ for units $i \neq j$ in the same strata and 0 for units from different strata, where $\rho_h = \sigma_{uh}^2 / (\sigma_{uh}^2 + \sigma_{eh}^2)$. For simplicity, it will be assumed that $\rho_h = \rho$ are equal for all strata.

Under model (2.1),

$$E_\xi[v_{hr}] = E_\xi[n_h^{-1} S_{hw}^2] = n_h^{-1} \sigma_h^2 (1-\rho)$$

where S_{hw}^2 is the within-stratum-h sample variance of $y_i - \boldsymbol{\beta}_U^{\mathrm{T}} \mathbf{x}_i$; and

$$\begin{aligned} E_\xi[B_h^2] &= E_\xi\left[\left(\bar{Y}_h - \hat{\bar{Y}}_{h(\text{syn})}\right)^2\right] \approx E_\xi\left[\left(\bar{Y}_h - \boldsymbol{\beta}^{\mathrm{T}} \bar{\mathbf{X}}_h\right)^2\right] \\ &= \text{var}_\xi[\bar{Y}_h] = \sigma_h^2 N_h^{-1} [1 + (N_h - 1)\rho]. \end{aligned}$$

To simplify expressions, we assume that n, N_h and H are all large, although we do not derive rigorous asymptotic results. Assuming that N_h is large, we firstly obtain $E_\xi[B_h^2] \approx \sigma_h^2 \rho$. Substituting for $E_\xi[v_{hr}]$ and $E_\xi[B_h^2]$ into (2.1) we get the anticipated MSE or approximate model assisted mean squared error, denoted AMSE_h :

$$\text{AMSE}_h = E_\xi \text{MSE}_p\left(\bar{y}_h^C; \bar{Y}_h\right) \approx (1-\phi_h)^2 n_h^{-1} \sigma_h^2 (1-\rho) + \phi_h^2 \sigma_h^2 \rho. \tag{2.3}$$

Optimizing with respect to ϕ_h we immediately obtain the optimal weight ϕ_h as:

$$\phi_{h(\text{opt})} = (1-\rho)[1+(n_h-1)\rho]^{-1}. \tag{2.4}$$

We substitute the optimum weight (2.4) into (2.3) to obtain the approximate optimum anticipated MSE:

$$\begin{aligned}
\text{AMSE}_h &= E_\xi \text{MSE}_p\left(\bar{y}_h^C[\phi_{h(\text{opt})}];\bar{Y}_h\right)\\
&\approx \left(n_h\rho[1+(n_h-1)\rho]^{-1}\right)^2 n_h^{-1}\sigma^2(1-\rho)+\left((1-\rho)[1+(n_h-1)\rho]^{-1}\right)^2\sigma^2\rho\\
&= \sigma_h^2\rho(1-\rho)[1+(n_h-1)\rho]^{-1}.
\end{aligned}$$

3 Optimizing the design

3.1 Optimal design for F

One way of measuring the performance of designs for small area estimation is with a linear combination of the anticipated MSEs of the small area mean and overall mean estimators. Following Longford (2006), but using anticipated MSEs instead of design-based MSEs, we define the criterion

$$\begin{aligned}
F &= \sum_{h\in U^q} N_h^q \text{AMSE}_h + \text{GN}_+^{(q)} E_\xi \text{var}_p\left[\hat{\bar{Y}}_r\right]\\
&= \sum_{h\in U^q} N_h^q \text{AMSE}_h + \text{GN}_+^{(q)} E_\xi \text{var}_p\left[\sum_{h\in U^q} P_h \bar{y}_{hr}\right]\\
&\approx \sum_{h\in U^q} N_h^q \text{AMSE}_h + \text{GN}_+^{(q)} E_\xi \sum_{h\in U^q} P_h^2 n_h^{-1} S_{hw}^2\\
&= \sum_{h\in U^q} N_h^q \sigma_h^2\rho(1-\rho)[1+(n_h-1)\rho]^{-1} + \text{GN}_+^{(q)} \sum_{h\in U^q} \sigma_h^2 P_h^2 n_h^{-1}(1-\rho)
\end{aligned} \tag{3.1}$$

where the weights N_h^q reflect the inferential priorities for area h, with $0 \le q \le 2$, and $N_+^{(q)} = \sum_{h\in U^q} N_h^q$, and $\bar{y}_{hr}$ is the grand mean estimator defined in Section 2. This objective reflects the fact that surveys have many stakeholders, some of whom will be only concerned with one specific small area, while others will place priority only on national estimators. Estimators for small regions are often considered a priority, particularly if they correspond to administrative or governmental jurisdictions, although smaller areas may be assigned less priority than larger regions. The quantity G is a relative priority coefficient. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small area estimation corresponds to large values of G, since when G is very large the second component in (3.1) dominates. The factor $N_+^{(q)}$ is introduced to appropriately scale for the effect of the absolute sizes of N_h^q and the number of areas on the relative priority G. The criterion in (3.1) is algebraically similar to the criterion in Longford (2006). Here, however, we adopt the model-assisted approach which treats the design-based inference as the real goal of survey sampling, but employs models to choose between valid randomization-based alternatives (e.g., Chapter 6 of Särndal, Swensson and Wretman 1992).

Suppose that national estimation has no priority $(G = 0)$, and the aim is to minimize (3.1) subject to a fixed total sampling cost function $C_f = \sum_h C_h n_h$, where C_h is the unit cost of surveying a unit in stratum h. The unique stationary point for this optimization is

$$n_{h,opt} = \frac{C_f \sqrt{N_h^q \sigma_h^2 C_h^{-1}}}{\sum_{h \in U^1} \sqrt{N_h^q \sigma_h^2 C_h}} + \frac{1-\rho}{\rho}\left(\frac{C \sqrt{N_h^q \sigma_h^2 C_h^{-1}}}{H^{-1} \sum_{h \in U^1} \sqrt{N_h^q \sigma_h^2 C_h}} - 1 \right) \tag{3.2}$$

where $C = H^{-1} \sum_h C_h$. We will concentrate on the case when unit costs are equal across strata, so that the constraint becomes $n = \sum_h n_h$ and (3.2) simplifies to

$$n_{h,opt} = \frac{n \sqrt{\sigma_h^2 N_h^q}}{\sum_{h \in U^1} \sqrt{\sigma_h^2 N_h^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_h^2 N_h^q}}{H^{-1} \sum_{h \in U^1} \sqrt{\sigma_h^2 N_h^q}} - 1 \right). \tag{3.3}$$

If there are other active constraints (e.g., minimum stratum sample sizes or maximum stratum MSEs), or if $G > 0$, then (3.2) and (3.3) do not apply and F must be minimized numerically, for example by NLP as in Choudhry et al. (2012).

In practice it would almost always be appropriate to set $0 \le q \le 2$, with $q = 0$ corresponding to all areas being equally important regardless of size, and $q = 2$ giving much greater weight to larger areas. (The value of $q = 2$ would lead to proportional allocation if direct estimators were used rather than composite - see for example Bankier 1988.) In many cases $q = 1$ would be a sensible compromise. For example, this has been used to motivate power allocations (Bankier 1988) for master household samples in Vietnam and South Africa (Kalton, Brick and Lê 2005, paragraph 76, page 89).

The first term in (3.3) is the optimal allocation for the direct estimator and corresponds to power allocation (Bankier 1988). The second term will be positive for more populous areas (large N_h) and negative for less populous areas. Therefore, the allocation optimal for model-assisted composite estimation has more dispersed subsample sizes $n_{h,opt}$ than the allocation that is optimal for direct estimators.

To understand the properties of the optimal allocation when $G > 0$, and to provide a non-iterative method, Molefe (2011, Chapter 3) derived Taylor Series approximations to the optimal n_h, based on small ρ. However, the resulting approximation tended to result in very large negative and very large positive values of $n_{h,opt}$ unless ρ is very small. (In practice, these would be truncated to either 0 or the population size, respectively.) Mathematically, the issue is apparently that the optimal n_h are quite nonlinear in ρ at $\rho = 0$, so that Taylor Series approximations are only a good approximation in a small neighbourhood of $\rho = 0$. Taylor Series based on small values of a function of both G and ρ were also considered but had similar difficulties, and so these approaches are not further discussed here.

3.2 Power allocation

Power allocations (Bankier 1988) are defined by

$$n_h = \frac{n N_h^p}{\sum_{h \in U^1} N_h^p} \tag{3.4}$$

for $h \in U^1$, where $0 \leq p \leq 1$. A special case is the square root allocation when $p = 1/2$. The exponent p is called the power of the allocation. Setting $p = 1$ results in proportional allocation and $p = 0$ results in equal allocation.

Bankier (1988) proposed choosing p based on perceived relative priorities. However, this was based on direct estimators being used in each stratum. We are interested in the case where composite estimation is to be used, and the objective is to obtain a low value for F in (3.1). We obtain numerically the value of p which minimizes F by one-dimensional optimization. We further consider imposing minimum stratum sample sizes, with p re-optimized accordingly. (Alternatively, maximum stratum MSE constraints could be imposed.)

4 Numerical study

We use data on the 26 cantons of Switzerland (Longford 2006); their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million. We assume that $n = 10{,}000$, $\rho = 0.025$ and $\sigma/\mu = 1$ (following Longford 2006). The last assumption only affects the magnitude of F and the relative root mean squared errors (RRMSE) but not the relativities across methods. It is satisfied if, for example, a prevalence rate of 50% is estimated. All calculations were performed in the R statistical environment (R Development Core Team 2012). Values of $q = 0, 0.5, 1, 1.5$ and 2, and values of $G = 0, 10$ and 100 were used, as in section 5.2 of Choudhry et al. (2012). The program used to produce all results is available in the appendix of Molefe and Clark (2014).

Six different allocations are evaluated in Tables 4.1-4.3. The value of F is shown for each design, relative to the value for equal allocation. Strata sample sizes were constrained in all allocations to lie between 1 and the population sizes, while still summing to n. The first design is equal allocation, then proportional allocation. The third design is the optimal design, which minimizes F in (3.1) by NLP subject to all stratum sample sizes being at least 1. The fourth design minimizes F subject to all stratum RRMSEs being 8% or less, which, from formula (3.1), is equivalent to a minimum stratum sample size of 113. For the third and fourth designs, NLP was carried out using the R package *Rsolnp* (Ghalanos and Theussl 2011). The fifth design is power allocation, where the exponent p is calculated to minimize F. The sixth design is power allocation with all stratum sample sizes constrained to be 113 or more, and with p calculated to minimize F reflecting these constraints. In both the fifth and sixth cases, p was calculated using the *optimize* function in R.

Table 4.1 shows the efficiency of the various methods when $G = 0$, where efficiency refers to the achieved values of F from formula (3.1), which is a weighted combination of MSEs of area composite estimators and an overall grand mean estimator. When $q = 0$, equal allocation is then optimal for F, and all of the allocation methods except proportional allocation return equal allocation. For larger values of q, Optimal for Composite is the most efficient, as expected. Imposing the area maximum RRMSE constraint of 8% increases F by 4% when $q = 2$, and has negligible effect (1.4% or less) for smaller q. The optimal power allocation has virtually identical efficiency to the optimal-for-composite allocation, both with and without the area RRMSE constraint. The unconstrained optimal-for-composite and power allocations are more efficient than proportional allocation when q is small, and about equally efficient for

$q \geq 1.5$. When the area RRMSE constraint is imposed, these designs suffer a small penalty, but are still more efficient than proportional except when $q = 2$.

Table 4.1
Relative efficiency of stratified designs for $G = 0$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	2.117	1.340	0.887	0.637	0.493
Optimal for composite	1.000	0.933	0.786	0.627	0.488
Optimal for composite with constraints	1.000	0.933	0.787	0.636	0.509
Optimal power allocation	1.000	0.933	0.786	0.628	0.490
Optimal power with constraints	1.000	0.933	0.787	0.636	0.509

Table 4.2 shows relative efficiencies for $G = 10$. As for when $G = 0$, the optimal-for-composite and optimal power designs perform very similarly, with a similar effect of imposing the area RRMSE constraint. The major difference compared to $G = 0$ is that proportional allocation is more efficient when G is larger. The optimal designs, even with the constraint imposed, remain more efficient than proportional allocation except for $q \geq 1.5$.

Table 4.2
Relative efficiency of stratified designs for $G = 10$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	1.360	0.944	0.701	0.568	0.491
Optimal for composite	0.875	0.784	0.668	0.565	0.490
Optimal for composite with constraints	0.875	0.784	0.670	0.575	0.505
Optimal power allocation	0.905	0.791	0.668	0.565	0.490
Optimal power with constraints	0.905	0.790	0.670	0.575	0.505

Table 4.3 shows efficiencies for large G ($G = 100$). Here, proportional allocation is close to the best design for all q. It is about equivalent to the unconstrained optimal designs for all $q \geq 0.5$, and more efficient than the constrained optimal designs for all $q \geq 1$. The relative performance of the four optimal designs is about the same as for $G = 0$ and $G = 10$.

Table 4.3
Relative efficiency of stratified designs for $G = 100$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	0.656	0.576	0.529	0.503	0.488
Optimal for composite	0.608	0.565	0.527	0.503	0.488
Optimal for composite with constraints	0.608	0.567	0.536	0.515	0.501
Optimal power allocation	0.624	0.567	0.528	0.503	0.488
Optimal power with constraints	0.612	0.568	0.536	0.515	0.501

Figure 4.1 shows the distribution of the area RRMSEs across the 26 cantons for $q \in \{0.5, 1, 1.5, 2\}$ when $G = 0$ for the four optimal designs. The results for $q = 0$ are not shown because the canton

sample sizes are then all equal for the optimal designs. The optimal for composite allocation (top left) shows a fairly tight range of area RRMSEs when $q = 0.5$, becoming more dispersed as q increases. The maximum RRMSEs are 6.6%, 9.4%, 13.8% and 15.6% for $q = 0.5$, 1, 1.5 and 2, respectively. Thus, for $q \geq 1$, some of the RRMSEs are undesirably large. The optimal for composite allocation with constraints forces all area RRMSEs to be 8% or less, shown by the top right panel. The bottom two panels show the corresponding optimal power allocations. The unconstrained power allocation is broadly similar to the unconstrained optimal for composite allocation, but less dispersed, with lower maximum area RRMSEs. The two constrained designs are very similar.

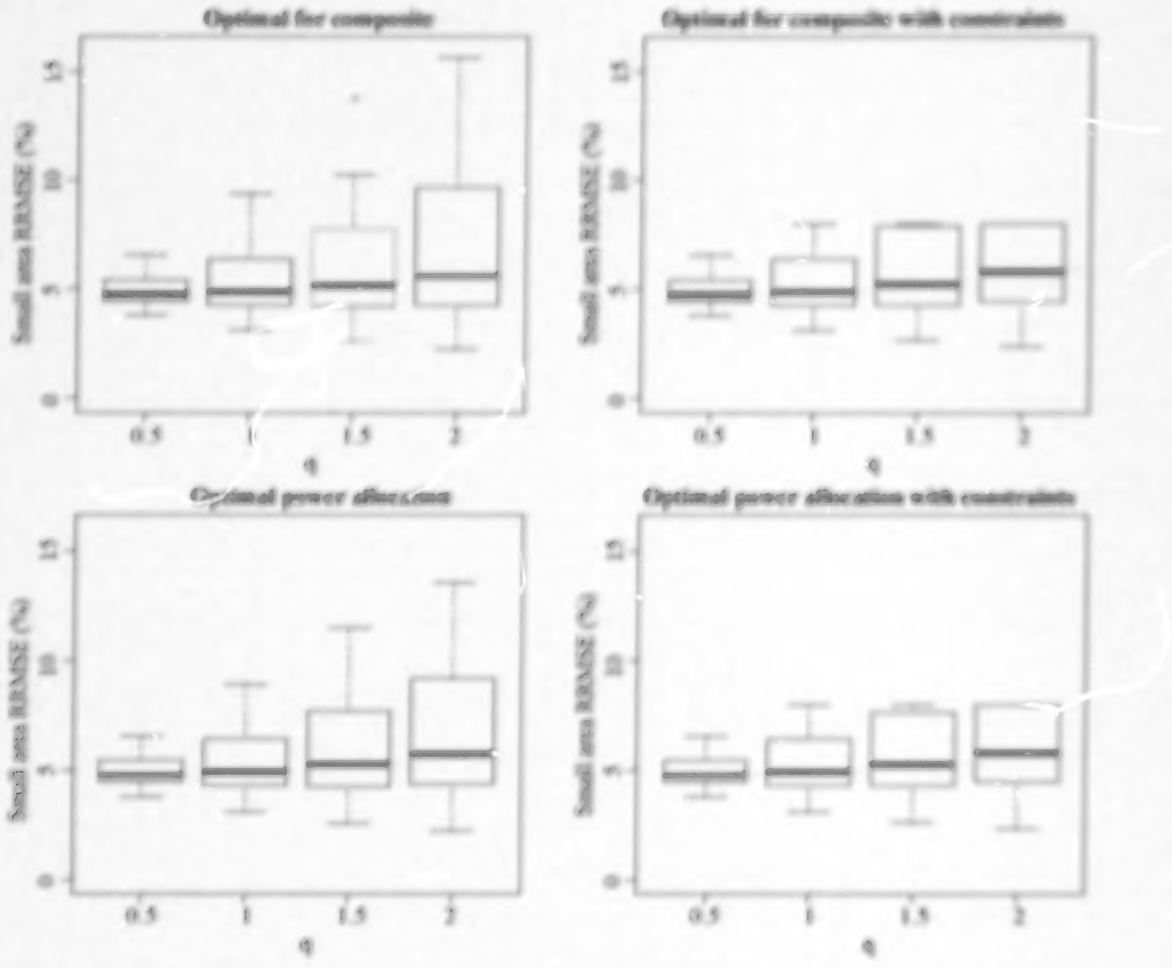

Figure 4.1 Distribution of anticipated relative root mean squared errors (RRMSE) (%) of estimated strata means for 4 allocations for various q with $G = 0$.

Table 4.4 shows the values of the optimal exponents calculated for the optimal power designs for each q and G. When G is 0 or 10, the optimal exponent p of the power allocation is very close to $q/2$, where q is the exponent in the definition of F in (3.1). For $G = 100$, the optimal exponent is quite close to 1, reflecting that for large G, F essentially reflects the variance of the grand mean, so that proportional allocation is nearly optimal. Table 4.5 shows the optimal power exponents when the area RRMSE constraints are applied. Applying these constraints has little effect on the optimal p.

Table 4.4
Optimal exponent in power allocation by G and q

	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
$G = 0$	0.000	0.277	0.537	0.837	1.111
$G = 10$	0.293	0.500	0.721	0.932	1.050
$G = 100$	0.750	0.852	0.956	0.985	1.008

Table 4.5
Optimal exponent in power allocation by G and q with constraint on strata RRMSEs

	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
$G = 0$	0.000	0.277	0.554	0.813	1.015
$G = 10$	0.293	0.513	0.729	0.898	1.036
$G = 100$	0.859	0.907	0.945	0.976	1.007

5 Conclusions

The anticipated MSE is a sensible objective criterion for sample design, because the particular sample which will be selected is not available in advance of the survey. Hence a criterion which averages over all possible samples is appropriate. Särndal et al. (1992, Chapter 12) base their optimal designs on the anticipated variance, which similarly averages over both model realizations and sample selection, although they consider only approximately design-unbiased estimators.

When both strata composite estimators and overall estimators are a priority, it makes sense to optimize an objective criterion which is a linear combination of the relevant anticipated MSE's. Allocations which are optimal in this sense give lower values of the objective function than either proportional or equal allocation. An optimal power allocation, $n_h \propto N_h^p$ where p is obtained numerically to minimize the objective function, is simpler and avoids the possibility of negative sample sizes which need to be truncated. Under conditions, it is very nearly as efficient as the optimal allocation. When there is no priority on national estimation ($G = 0$), the optimal exponent turns out to be close to $p = q/2$, where q is the exponent applied to stratum population sizes in the objective criterion. This removes the need to perform an optimization. Thus, we recommend an objective criterion very similar to that of Longford (2006), but we suggest a simple power allocation with $p = q/2$ when $G = 0$, rather than the optimal allocation for F. This extends the the domain of application of power allocation to surveys using stratum composite estimators.

Rather than just relying on the overall objective criterion to appropriately balance resources across strata, it may often be desirable to also impose minimum stratum sample sizes or maximum stratum RRMSEs. These were successfully implemented using NLP. In the Swiss canton example in Section 4, an upper limit of 8% for stratum RRMSEs significantly reduced the highest RRMSE with little loss in the objective criterion. More complex constraints, for example on cross-strata domains or for multiple variables of interest, could also be implemented using NLP.

Acknowledgements

The authors wish to thank Professors Raymond Chambers and David Steel for their helpful suggestions on this paper.

Appendix

Derivation of (3.2)

The steps of this derivation are similar to Longford (1999) although F is defined differently and unequal costs are allowed. A stationary point of (3.1) subject to $C_T = \sum_{h=1}^{H} C_h n_h$ is given by

$$0 = \frac{\partial F}{\partial n_h} + \lambda C_h$$

$$= -N_h^2 w_h^2 \sigma_h^2 (1-\rho)\{(1-\rho)n_h + (n_h - 1)\rho\}^{-2} + \lambda C_h$$

Writing $\gamma = \lambda\rho^{-2}(1-\rho)^{-2}$ and rearranging gives

$$\begin{aligned} (1-\rho)\{(1-\rho)n_h + (n_h-1)\rho\}^{-2} &= \rho^2 C_h N_h^2 w_h^2 \sigma_h^2 \\ (1-\rho)n_h + (n_h-1)\rho &= \gamma^{-1/2}\sqrt{C_h N_h^2 w_h^2 \sigma_h^2} \\ n_h &= \gamma^{-1/2}\rho^{-1}\sqrt{C_h N_h^2 w_h^2 \sigma_h^2} - \frac{1-\rho}{\rho} \end{aligned} \tag{A.1}$$

Substituting into the constraint $C_T = \sum_{h=1}^{H} C_h n_h$ and solving for γ gives

$$\gamma^{-1/2} = \frac{C_T\rho + (1-\rho)\sum_{h=1}^{H} C_h}{\sum_{h=1}^{H}\sqrt{C_h N_h^2 w_h^2 \sigma_h^2}}$$

where $C_T = H^{-1}\sum_{h=1}^{H} C_h$. Substituting back into (A.1) and rearranging gives the result.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Chaudhuri, S.H., Ren, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-33.

Ghalanos, A., and Theussl, S. (2011). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. http://cran.r-project.org/package=Rsolnp, R package version 1.11.

Hidiroglou, M.A., and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 1, 67-78.

Kalton, G., Brick, J. and Lê, T. (2005). *Household Sample Surveys in Developing and Transition Countries*. United Nations: Statistics Division, Department of Economic and Social Affairs, ser. M no. Series F. http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf [accessed 24-May-2013].

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 1, 87-96.

Marker, D.A. (1991). Producing small area estimates from National Surveys: Methods of minimizing use of indirect estimators. *Survey Methodology*, 17, 2, 183-196.

Molefe, W., and Clark, R.G. (2014). Model-assisted optimal allocation for planned domains using composite estimation. http://eprints.qut.edu.au/publications/EXPR/47615.html, Statistics Working Paper 09-14.

Molefe, W.B. (2011). Sample Design for Small Area Estimation. Ph.D. thesis, University of Wollongong, http://ro.uow.edu.au/theses/3491.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. http://www.R-project.org/, ISBN 3-900051-07-0.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.